1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] There are no societal impact as this is a purely theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Further Experimental Results and Details

Run-time specification. The experiments in the main text and in the appendix were run on a normal PC laptop with Processor Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz, 2112 Mhz, 4 Core(s), 8 Logical Processor(s), 16GB RAM. It took around 1.5 hours to run all the experiments.

A.1 Break-down Point

As expected, we do verify that as the sub-sample size becomes $m = O(n)$, where the estimate is both not stable and also does not optimize over hypothesis spaces with small critical radius or that satisfy the Donsker property, then the estimate without cross-fitting breaks down, while the estimate with cross-fitting maintains a decent performance, despite the high-variance of the nuisance estimate.

		bias	std	std_est	cov95
n=50, n_x=1	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.010	0.200	0.135	0.805
n=50, n_x=2	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.017	0.204	0.138	0.795
n=100, n_x=1	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.003	0.148	0.097	0.794
n=100, n_x=2	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.008	0.146	0.098	0.797
n=500, n_x=1	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.002	0.066	0.045	0.798
n=500, n_x=2	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.002	0.064	0.045	0.838
n=1000, n_x=1	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.000	0.046	0.031	0.817
n=1000, n_x=2	cv=1	0.500	0.000	0.000	0.000
	cv=2	0.001	0.045	0.032	0.829

(a) Sub-sampled 1-NN with $m = n$

Figure 2: Comparison of bias, variance and coverage properties, with (cv=2) and without (cv=1) cross-fitting (sample splitting), for the estimation of the treatment effect in the partially linear model, when a sub-sampled 1-NN estimation is used for the nuisance function estimation. n is the number of samples and n_x the number of controls.

A.2 Random Forest Experiments

		bias	std	std_est	cov95
n=50, n_x=5	cv=1	0.102	0.149	0.144	0.873
	cv=2	0.103	0.165	0.143	0.836
n=50, n_x=10	cv=1	0.098	0.136	0.134	0.861
	cv=2	0.099	0.148	0.133	0.846
n=100, n_x=5	cv=1	0.066	0.102	0.101	0.894
	cv=2	0.064	0.109	0.101	0.877
n=100, n_x=10	cv=1	0.074	0.099	0.097	0.873
	cv=2	0.072	0.106	0.097	0.846
n=500, n_x=5	cv=1	0.020	0.044	0.045	0.930
	cv=2	0.021	0.046	0.045	0.919
n=500, n_x=10	cv=1	0.027	0.045	0.044	0.908
	cv=2	0.026	0.046	0.044	0.896
n=1000, n_x=5	cv=1	0.012	0.031	0.032	0.923
	cv=2	0.013	0.032	0.032	0.925
n=1000, n_x=10	cv=1	0.015	0.033	0.032	0.910
	cv=2	0.016	0.034	0.032	0.896

(a) Sub-sampled Random Forest with $m = n^{0.49}$

		bias	std	std_est	cov95
n=50, n_x=5	cv=1	0.022	0.176	0.148	0.896
	cv=2	0.011	0.191	0.144	0.845
n=50, n_x=10	cv=1	0.026	0.167	0.145	0.904
	cv=2	0.002	0.180	0.140	0.869
n=100, n_x=5	cv=1	0.016	0.116	0.103	0.909
	cv=2	0.013	0.125	0.101	0.877
n=100, n_x=10	cv=1	0.018	0.114	0.103	0.908
	cv=2	0.016	0.128	0.100	0.870
n=500, n_x=5	cv=1	0.012	0.050	0.046	0.908
	cv=2	0.008	0.053	0.045	0.900
n=500, n_x=10	cv=1	0.013	0.048	0.046	0.923
	cv=2	0.011	0.051	0.045	0.911
n=1000, n_x=5	cv=1	0.010	0.035	0.033	0.924
	cv=2	0.006	0.036	0.032	0.912
n=1000, n_x=10	cv=1	0.012	0.035	0.033	0.909
	cv=2	0.008	0.038	0.032	0.893

(b) Sub-sampled Random Forest with $m = n^{10/11}$

Figure 3: Comparison of bias, variance and coverage properties, with (cv=2) and without (cv=1) cross-fitting (sample splitting), for the estimation of the treatment effect in the partially linear model, when a sub-sampled Random Forest estimation is used for the nuisance function estimation. n is the number of samples and n_x the number of controls.

A.3 Quantile-Quantile Plots

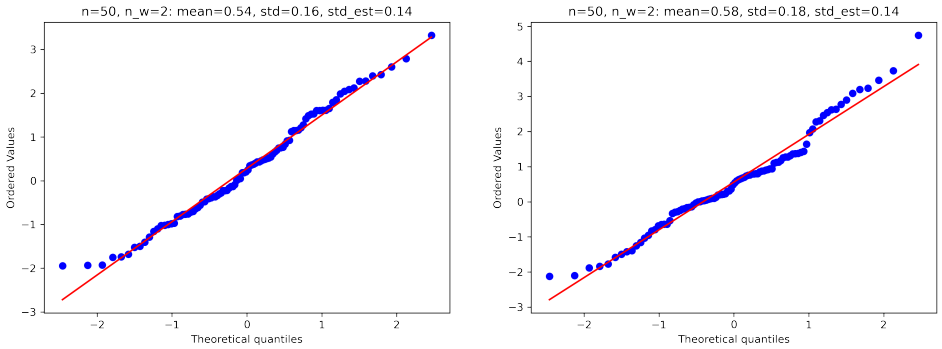

Figure 4: Quantile-Quantile (Q-Q) plots for sub-sampled 1-NN, with (right) and without (left) cross-fitting.

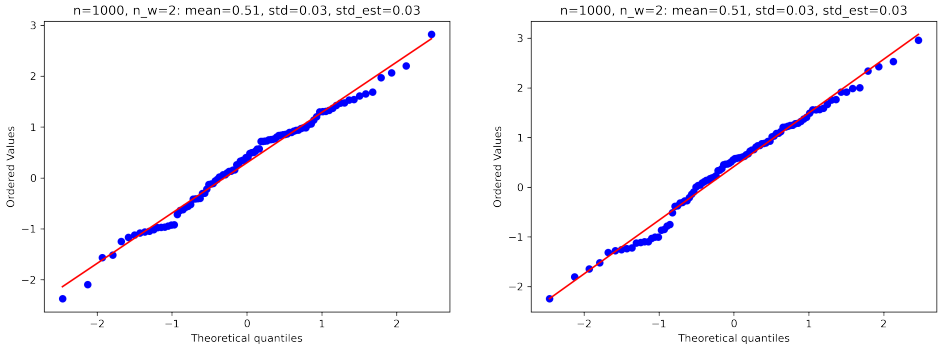

Figure 5: Quantile-Quantile (Q-Q) plots for sub-sampled 1-NN, with (right) and without (left) cross-fitting.

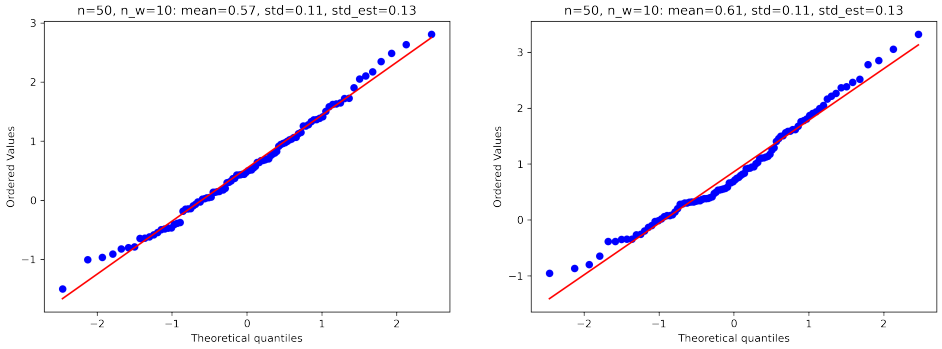

Figure 6: Quantile-Quantile (Q-Q) plots for sub-sampled Random Forest, with (right) and without (left) cross-fitting.

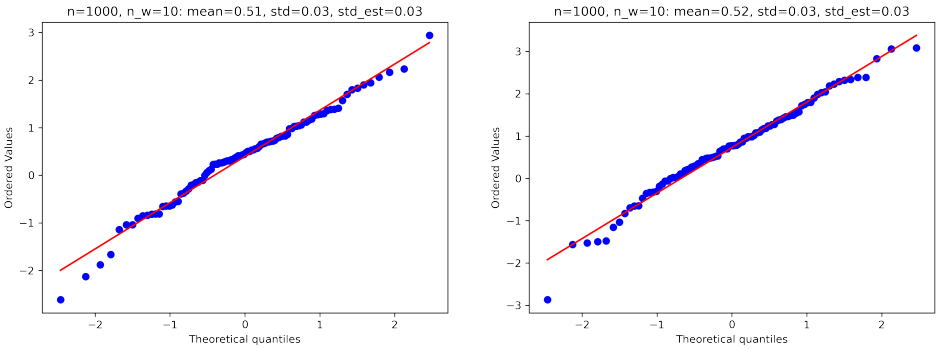

Figure 7: Quantile-Quantile (Q-Q) plots for sub-sampled Random Forest, with (right) and without (left) cross-fitting.

B Proof of Theorem 1

Proof. For any $g \in \mathcal{G}$, by the linearity of the moment with respect to θ:

$$\begin{aligned} A(g)\left(\hat{\theta}-\theta_0\right) &= M(\hat{\theta}, g) - M(\theta_0, g) \\ &= M(\hat{\theta}, g) - M_n(\hat{\theta}, g) + M(\theta_0, g_0) - M(\theta_0, g) + M_n(\hat{\theta}, g). \end{aligned}$$

Moreover, for any g, with $\|g - g_0\|_2 = o_p(1)$:

$$\begin{aligned} A(g)\left(\hat{\theta}-\theta_0\right) &= A(g_0)\left(\hat{\theta}-\theta_0\right) + (A(g) - A(g_0))\left(\hat{\theta}-\theta_0\right) \\ &= A(g_0)\left(\hat{\theta}-\theta_0\right) + O\left(\|g-g_0\|_2 \|\hat{\theta}-\theta_0\|_2\right) \\ &= A(g_0)\left(\hat{\theta}-\theta_0\right) + o_p\left(\|\hat{\theta}-\theta_0\|_2\right). \end{aligned}$$

Thus for any g, with $\|g - g_0\|_2 = o_p(1)$:

$$A(g_0)\left(\hat{\theta}-\theta_0\right) = M(\hat{\theta}, g) - M_n(\hat{\theta}, g) + M(\theta_0, g_0) - M(\theta_0, g) + M_n(\hat{\theta}, g) + o_p(\|\hat{\theta}-\theta_0\|_2).$$

Let $G_n(\theta, g) := M(\theta, g) - M_n(\theta, g)$, then we have:

$$A(g_0)\left(\hat{\theta}-\theta_0\right) = G_n(\hat{\theta}, g) + M(\theta_0, g_0) - M(\theta_0, g) + M_n(\hat{\theta}, g) + o_p(\|\hat{\theta}-\theta_0\|_2).$$

Applying the above for $g = \hat{g}$ and by the definition of $\hat{\theta}$, we have:

$$\begin{aligned} A(g_0)\left(\hat{\theta}-\theta_0\right) &= G_n(\hat{\theta}, \hat{g}) + M(\theta_0, g_0) - M(\theta_0, \hat{g}) + M_n(\hat{\theta}, \hat{g}) + o_p(\|\hat{\theta}-\theta_0\|_2) \\ &= G_n(\hat{\theta}, \hat{g}) + M(\theta_0, g_0) - M(\theta_0, \hat{g}) + o_p(n^{-1/2} + \|\hat{\theta}-\theta_0\|_2). \end{aligned}$$

Applying Neyman orthogonality and bounded second derivative of the moment with respect to g:

$$M(\theta_0, g_0) - M(\theta_0, g) = D_g\, M(\theta_0, g_0)[g_0 - g] + O\left(\|g-g_0\|_2^2\right) = O\left(\|g-g_0\|_2^2\right) = o_p(n^{-1/2}).$$

Thus we have that:

$$A(g_0)\left(\hat{\theta}-\theta_0\right) = G_n(\hat{\theta}, \hat{g}) + o_p(n^{-1/2} + \|\hat{\theta}-\theta_0\|_2).$$

Now we decompose the empirical process part into an asymptotically normal component and asymptotically equicontinuous parts that converge to zero in probability:

$$G_n(\hat{\theta}, \hat{g}) = G_n(\theta_0, g_0) + \left(G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g})\right) + \left(G_n(\theta_0, \hat{g}) - G_n(\theta_0, g_0)\right).$$

By the linearity of the moment, the middle term can be written as:

$$G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g}) = \ (A(\hat{g}) - A_n(\hat{g}))' (\hat{\theta} - \theta_0).$$

Note that by a triangle inequality:

$$\|A(\hat{g}) - A_n(\hat{g})\|_{op} \leq \ \|A(g_0) - A_n(g_0)\|_{op} + \|A(\hat{g}) - A(g_0) - (A_n(\hat{g}) - A_n(g_0))\|_{op}$$

Note that the first quantity is a simple centered empirical process and hence assuming that $a_{i,j}(Z; g_0)$ has bounded variance, by classical results in empirical process theory we have that:

$$\|A(g_0) - A_n(g_0)\|_{op} = o_p(1)$$

Moreover, by our stochastic equicontinuity condition we have that:

$$\|A(\hat{g}) - A(g_0) - (A_n(\hat{g}) - A_n(g_0))\|_{op} = o_p(n^{-1/2}) = o_p(1).$$

Thus we get that $\|A(\hat{g}) - A_n(\hat{g})\|_{op} = o_p(1)$, and therefore:

$$G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g}) = \ o_p\left(\|\hat{\theta} - \theta_0\|_2\right).$$

Moreover, since by our stochastic equicontinuity conditions:

$$\sqrt{n}\,\|A(\hat{g}) - A(g_0) - (A_n(\hat{g}) - A_n(g_0))\|_{op} = o_p(1)$$
$$\sqrt{n}\,\|V(\hat{g}) - V(g_0) - (V_n(\hat{g}) - V_n(g_0))\|_2 = o_p(1)$$

we have by triangle inequality, the definition of the operator norm, and the fact that $\|\theta_0\|_2 = O(1)$ that:

$$\begin{aligned}\|G_n(\theta_0, \hat{g}) - G_n(\theta_0, g_0)\|_2 \leq \ & \|A(\hat{g}) - A(g_0) - (A_n(\hat{g}) - A_n(g_0))\|_{op}\ \|\theta_0\|_2 \\ & + \|V(\hat{g}) - V(g_0) - (V_n(\hat{g}) - V_n(g_0))\|_2 \\ = \ & o_p(n^{-1/2}).\end{aligned}$$

Thus we can conclude that:

$$A(g_0)\left(\hat{\theta} - \theta_0\right) = \ G_n(\theta_0, g_0) + o_p(n^{-1/2} + \|\hat{\theta} - \theta_0\|_2).$$

Assuming that the inverse $A(g_0)^{-1}$ exists, we can re-arrange to:

$$\hat{\theta} - \theta_0 = A(g_0)^{-1} G_n(\theta_0, g_0) + o_p\left(n^{-1/2} + \|\hat{\theta} - \theta_0\|_2\right).$$

Since $G_n(\theta_0, g_0)$ is a mean-zero empirical process, we have that $\|G_n(\theta_0, g_0)\|_2 = O_p(n^{-1/2})$. Thus the above equation implies that $\|\hat{\theta} - \theta_0\|_2 = O_p(n^{-1/2})$. Thus we get:

$$\hat{\theta} - \theta_0 = A(g_0)^{-1} G_n(\theta_0, g_0) + o_p\left(n^{-1/2}\right)$$

or equivalently that:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \sqrt{n} A(g_0)^{-1} G_n(\theta_0, g_0) + o_p\left(1\right).$$

The first term converges in distribution to the claimed normal limit by invoking the Central Limit Theorem. Thus the theorem follows by Slutsky's theorem. □

C Proof of Lemma 3

Proof. Before diving into the proof, recall that $c(Z_{1:n}, x) = \arg\min_{i \leq n} |X_i - x|$, and define:

$$c_1^* := c(Z_{1:n}, \frac{1}{2}),$$
$$c_2^* := \begin{cases} \underset{i \leq n \text{ s.t } X_i \leq \frac{1}{2}}{\arg\min} |\frac{1}{2} - X_i| & \text{if } X_{c_1^*} > \frac{1}{2}, \\ \underset{i \leq n \text{ s.t } X_i > \frac{1}{2}}{\arg\min} |\frac{1}{2} - X_i| & \text{if } X_{c_1^*} \leq \frac{1}{2}. \end{cases}.$$

That is, we let c_1^* be the index of the nearest example in $\{X_1, \ldots, X_n\}$ to $\frac{1}{2}$, and let c_2^* be the index of the nearest example to $\frac{1}{2}$ on the other side of $\frac{1}{2}$ from.

A new observation $Z = (X, Y)$, where $X \sim \text{unif}[0, 1]$ and $Y = \mathbb{I}(X \leq 0.5)$, will be misclassified if $Y_{c(Z_{1:n}, X)}$ is different from Y. Therefore it is mislabeled if it falls in the following set:

$$\mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big) := \begin{cases} \Big[\frac{1}{2}, \frac{1}{2}\big(X_{c_1^*} + X_{c_2^*}\big)\Big] & \text{if } X_{c_1^*} \leq \frac{1}{2}, \\ \Big[\frac{1}{2}\big(X_{c_1^*} + X_{c_2^*}\big), \frac{1}{2}\Big] & \text{otherwise.} \end{cases}$$

For a given pair of random variables X_{i_1}, X_{i_2}, we write:

$$\lambda_{X_{i_1,i_2}} := n\mathbb{P}\big(X \in \mathcal{E}(X_{i_1}, X_{i_2}) \mid X_{i_1}, X_{i_2}\big).$$

We note $B_1 := \{i \leq n | X_i \leq 1/2\}$.

Now we remark that if $X_{c_1^*} \leq 0.5$ then $\Big|1 - \Big[X_{c_1^*} + X_{c_2^*}\Big]\Big| = X_{c_1^*} + X_{c_2^*} - 1$ and

$$X_{c_2^*} - (1 - X_{c_1^*}) | X_{c_1^*} \sim \min_{i \in [n] \setminus B_1} U_i(X_{c_1^*})$$

where $U_i(X_{c_1^*}) | X_{c_1^*} \sim_{i.i.d} \text{unif}[0, \text{X}_{\text{c}_1^*}]$. Here $\sim$ means "has the same distribution as."

Therefore we have

$$\begin{aligned} \mathbb{P}\Big(X_{c_1^*} + X_{c_2^*} - 1 \leq \frac{2t}{n} \Big| X_{c_1^*} \leq 0.5, X_{c_1^*}\Big) &= \mathbb{P}\Big(\min_{i \in [n] \setminus B_1} U_i(X_{c_1^*}) \leq \frac{2t}{n} \Big| X_{c_1^*} \leq 0.5, X_{c_1^*}\Big) \\ &= 1 - \Big(1 - \frac{2t}{nX_{c_1^*}}\Big)^{n - |B_1|}, \end{aligned}$$

where $|B_1|$ denotes the cardinality of set B_1.

Therefore as $X_{c_1^*} \to 0.5$ and $|B_1| \to n/2$ we have

$$P\Big(X_{c_1^*} + X_{c_2^*} - 1 \leq \frac{2t}{n} \Big| X_{c_1^*} \leq 0.5, X_{c_1^*}\Big) \to 1 - e^{-2t}.$$

Similarly we remark that if $X_{c_1^*} > 0.5$ then $\Big|1 - \Big[X_{c_1^*} + X_{c_2^*}\Big]\Big| = 1 - X_{c_1^*} - X_{c_2^*}$ and

$$1 - X_{c_1^*} - X_{c_2^*} | X_{c_1^*} \sim \min_{i \in B_1} U_i^{bis}(X_{c_1^*})$$

where $U_i^{bis}(X_{c_1^*}) | X_{c_1^*} \sim_{i.i.d} \text{unif}[0, 1 - \text{X}_{\text{c}_1^*}]$.

Therefore we have

$$\begin{aligned} \mathbb{P}\Big(1 - X_{c_1^*} - X_{c_2^*} \leq \frac{2t}{n} \Big| X_{c_1^*} > 0.5, X_{c_1^*}\Big) &= \mathbb{P}\Big(\min_{i \in B_1} U_i^{bis}(X_{c_1^*}) \leq \frac{2t}{n} \Big| X_{c_1^*} > 0.5, X_{c_1^*}\Big) \\ &= 1 - \Big(1 - \frac{2t}{n(1 - X_{c_1^*})}\Big)^{|B_1|}. \end{aligned}$$

Therefore as $X_{c_1^*} \to 0.5$ and $|B_1| \to n/2$ we have

$$P\Big(1 - X_{c_1^*} - X_{c_2^*} \leq \frac{2t}{n} \Big| X_{c_1^*} > 0.5, X_{c_1^*}\Big) \to 1 - e^{-2t}.$$

This directly implies that

$$\lambda_{X_{c_1^*, c_2^*}} = \frac{n}{2}\Big|1 - \Big[X_{c_1^*} + X_{c_2^*}\Big]\Big| \xrightarrow{d} \text{Exp}(2),$$

where $\text{Exp}(2)$ denotes an exponential distribution with rate parameter 2.

Moreover, we now also show that the expectation $\mathbb{E}\Big[\lambda_{X_{c_1^*, c_2^*}}\Big] = O(1)$.

As a first step, we note that we can write

$$\begin{aligned}
&\mathbb{E}\left[\lambda_{X_{c_1^*,c_2^*}}\right] = n\mathbb{P}\left(X \in \mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)\right) \\
&\leq n\mathbb{P}\left(X \in \mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big), \left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right) + n\mathbb{P}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| > \frac{1}{4}\right).
\end{aligned}$$

By Azuma's concentration inequality, we know that

$$n\mathbb{P}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| > \frac{1}{4}\right) \leq 2ne^{-\frac{32}{n}} \to 0 \text{ as } n \to \infty.$$

To treat the other term in the sum, we have that

$$\begin{aligned}
&n\mathbb{P}\left(X \in \mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big), \left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right) \\
&= n\mathbb{E}\left[\mathbb{P}\left(X \in \mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big), \left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\middle| X_{1:n}\right)\right] \\
&= n\mathbb{E}\left[\mathbb{P}\left(X \in \mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)\middle| X_{1:n}\right)\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&= n\mathbb{E}\left[\left|\mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)\right|\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right]
\end{aligned}$$

where $X_{1:n} := (X_1, \ldots, X_n)$ and $\left|\mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)\right|$ denotes the length of the interval $\mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)$.

Now by triangle inequality

$$\begin{aligned}
&n\mathbb{E}\left[\left|\mathcal{E}\Big(X_{c_1^*}, X_{c_2^*}\Big)\right|\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&= n\mathbb{E}\left[\frac{1}{2}\left|1 - \left(X_{c_1^*} + X_{c_2^*}\right)\right|\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&\leq \frac{n}{2}\mathbb{E}\left[\left|\frac{1}{2} - X_{c_1^*}\right|\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] + \frac{n}{2}\mathbb{E}\left[\left|\frac{1}{2} - X_{c_2^*}\right|\mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&= \frac{n}{2}\mathbb{E}\left[\min_{i\in B_1} U_i \cdot \mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] + \frac{n}{2}\mathbb{E}\left[\min_{i\in[n]\setminus B_1} U_i \cdot \mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&= n\mathbb{E}\left[\min_{i\in B_1} U_i \cdot \mathbb{I}\left(\left|\frac{|B_1|}{n} - \frac{1}{2}\right| \leq \frac{1}{4}\right)\right] \\
&\leq n\mathbb{E}\left[\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right)\right].
\end{aligned}$$

where $U_i := \left|\frac{1}{2} - X_i\right| \sim_{\text{i.i.d.}} \text{unif}[0, 0.5]$, and the penultimate line follows from symmetry.
We can express the expectation in terms of integrals of tail probabilities as

$$\begin{aligned}
&n\mathbb{E}\left[\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right)\right] = n\mathbb{E}\left[\mathbb{E}\left[\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \,\middle|\, |B_1|\right]\right] \\
&= n\mathbb{E}\left[\int_0^\infty \mathbb{P}\left(\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \geq t \,\middle|\, |B_1|\right) dt\right] \\
&= n\mathbb{E}\left[\frac{1}{|B_1|}\int_0^\infty \mathbb{P}\left(\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \geq \frac{t}{|B_1|} \,\middle|\, |B_1|\right) dt\right].
\end{aligned}$$

Here

$$\begin{aligned}
&\mathbb{P}\left(\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \geq \frac{t}{|B_1|} \,\middle|\, |B_1|\right) \\
&= \left(1 - \frac{2t}{|B_1|}\right)^{|B_1|} \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \leq e^{-2t} \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right)
\end{aligned}$$

where we have used the inequality that $1 - x \leq e^{-x}$ for all x.

Hence, we have

$$
\begin{aligned}
&n\mathbb{E}\left[\frac{1}{|B_1|}\int_0^\infty \mathbb{P}\left(\min_{i\in B_1} U_i \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) \geq \frac{t}{|B_1|} \;\middle|\; |B_1|\right) dt\right] \\
&\leq n\mathbb{E}\left[\frac{1}{|B_1|}\int_0^\infty e^{-2t} \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right) dt\right] = \frac{n}{2}\mathbb{E}\left[\frac{1}{|B_1|} \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right)\right] \\
&\leq \frac{n}{2}\mathbb{E}\left[\frac{4}{n} \cdot \mathbb{I}\left(|B_1| \geq \frac{n}{4}\right)\right] = 2\mathbb{P}\left(|B_1| \geq \frac{n}{4}\right) \leq 2.
\end{aligned}
$$

Altogether, we have shown that

$$
\mathbb{E}\left[\lambda_{X_{c_1^*, c_2^*}}\right] = O(1). \tag{2}
$$

The key point is to note that

$$
\begin{aligned}
\|\hat{g}(Z_{1:n})(Z)\|_2^2 &= n^{1/3}\mathbb{P}(Y \neq Y_{c(Z_{1:n}, X)}) \\
&\leq n^{1/3}\mathbb{P}(X \in \mathcal{E}(X_{c_1^*}, X_{c_2^*})) \xrightarrow{(a)} 0.
\end{aligned}
$$

where (a) comes from realizing that

$$
\mathbb{P}(X \in \mathcal{E}(X_{c_1^*}, X_{c_2^*})) = \frac{1}{n}\mathbb{E}\left[\lambda_{X_{c_1^*, c_2^*}}\right].
$$

Therefore we proved the first point. Moreover if we denote

$$
\begin{aligned}
c_1^{*(-1)} &:= c(Z_{1:n}^{(-1)}, \frac{1}{2}), \\
c_2^{*(-1)} &:= \begin{cases} \underset{i \leq n \text{ s.t } X_i^{(-1)} \leq \frac{1}{2}}{\arg\min} |\frac{1}{2} - X_i^{(-1)}| & \text{if } X^{(-1)}_{c_1^{*(-1)}} > \frac{1}{2}, \\ \underset{i \leq n \text{ s.t } X_i^{(-1)} > \frac{1}{2}}{\arg\min} |\frac{1}{2} - X_i^{(-1)}| & \text{if } X^{(-1)}_{c_1^{*(-1)}} \leq \frac{1}{2} \end{cases},
\end{aligned}
$$

where $Z_{1:n}^{(-1)}$ is $Z_{1:n}$ with the first observation replaced with an independent copy $\tilde{Z}_1 = (\tilde{X}_1, \tilde{Y}_1)$, then we remark that

$$
\begin{aligned}
&\mathbb{P}\big(c_1^{*(-1)} \neq c_1^* \text{ or } \mathrm{c}_2^{*(-1)} \neq \mathrm{c}_2^*\big) \\
&\leq \mathbb{P}\big(c_1^{*(-1)} \neq c_1^*\big) + \mathbb{P}(c_2^{*(-1)} \neq c_2^*) \\
&\leq \mathbb{P}(1 = c_1^*) + \mathbb{P}(1 = c_1^{*(-1)}) + \mathbb{P}(1 = c_2^*) + \mathbb{P}(1 = c_2^{*(-1)}) \overset{(b)}{\leq} \frac{4}{n}
\end{aligned}
$$

where (b) comes from symmetry (each observation has an equal chance of being c_1^*, for example). Moreover we note that if neither $c_1^{*(-1)} \neq c_1^*$ nor $\mathrm{c}_2^{*(-1)} \neq \mathrm{c}_2^*$ then we have $\mathcal{E}(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}) = \mathcal{E}(X_{c_1^*}, X_{c_2^*})$.

Now for ease of notation denote

$$
X_{c_1^*, c_2^*} := \left(X_{c_1^*}, X_{c_2^*}\right),
$$

$$
X_{c_1^{*(-1)}, c_2^{*(-1)}} := \left(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\right),
$$

$$
X_{c_1^*, c_2^*, c_1^{*(-1)}, c_2^{*(-1)}} := \left(X_{c_1^*}, X_{c_2^*}, X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\right),
$$

and

$$
\mathcal{E}\left(X_{c_1^*, c_2^*, c_1^{*(-1)}, c_2^{*(-1)}}\right) := \mathcal{E}(X_{c_1^*}, X_{c_2^*}) \triangle \mathcal{E}\Big(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\Big).
$$

Then we have that there is $C < \infty$ such that

$$\|\nu(Z,\hat{g}) - \nu(Z,\hat{g}^{(-1)})\|_2 = \sqrt{n}\sqrt{\mathbb{P}\Big(X \in \mathcal{E}\left(X_{c_1^*,c_2^*,c_1^{*(-1)},c_2^{*(-1)}}\right)\Big)}$$

$$= \sqrt{n\mathbb{E}\left[\mathbb{P}\left(X \in \mathcal{E}\left(X_{c_1^*,c_2^*,c_1^{*(-1)},c_2^{*(-1)}}\right)\Big| X_{c_1^*,c_2^*,c_1^{*(-1)},c_2^{*(-1)}}\right)\mathbb{I}\left(c_1^{*(-1)} \neq c_1^* \text{ or } c_2^{*(-1)} \neq c_2^*\right)\right]}$$

$$\leq \sqrt{n\mathbb{E}\left[\left(\left|\mathcal{E}(X_{c_1^*,c_2^*})\right| + \left|\mathcal{E}\left(X_{c_1^{*(-1)},c_2^{*(-1)}}\right)\right|\right)\mathbb{I}\left(c_1^{*(-1)} \neq c_1^* \text{ or } c_2^{*(-1)} \neq c_2^*\right)\right]}$$

$$\leq \sqrt{2n\mathbb{E}\left[\left|\mathcal{E}(X_{c_1^*,c_2^*})\right|\mathbb{I}\left(c_1^{*(-1)} \neq c_1^* \text{ or } c_2^{*(-1)} \neq c_2^*\right)\right]} \text{ by symmetry}$$

$$\leq \sqrt{2n\mathbb{E}\left[\left|\mathcal{E}(X_{c_1^*,c_2^*})\right|\left(\mathbb{I}\left(c_1^* = 1\right) + \mathbb{I}\left(c_1^{*(-1)} = 1\right) + \mathbb{I}\left(c_2^* = 1\right) + \mathbb{I}\left(c_2^{*(-1)} = 1\right)\right)\right]}$$

$$\overset{(c)}{=} \sqrt{2n\mathbb{P}\big(X \in \mathcal{E}(X_{c_1^*,c_2^*})\big)\left(\mathbb{P}\left(c_1^* = 1\right) + \mathbb{P}\left(c_1^{*(-1)} = 1\right) + \mathbb{P}\left(c_2^* = 1\right) + \mathbb{P}\left(c_2^{*(-1)} = 1\right)\right)}$$

$$= \frac{\sqrt{8}}{\sqrt{n}}\sqrt{\mathbb{E}\left[\lambda_{X_{c_1^*,c_2^*}}\right]} \overset{(d)}{\leq} \frac{C}{\sqrt{n}}$$

where to get (c) we exploited independence of $(X_{c_1^*}, X_{c_2^*})$ and the events $\{c_1^* = 1\}$, $\{c_2^* = 1\}$, $\{c_1^{*(-1)} = 1\}$, $\{c_2^{*(-1)} = 1\}$ and where to get (d) we exploited (2).

Moreover we also notice that

$$\begin{aligned}
&\mathbb{P}\left(c_1^{*(-1)} \neq c_1^*\right) \\
&\geq \mathbb{P}\left(c_1^* = 1, \tilde{X}_1 \notin [\frac{1}{2} - \left|\frac{1}{2} - X_{c_3^*}\right|, \frac{1}{2} + \left|\frac{1}{2} - X_{c_3^*}\right|]\right) \\
&= \mathbb{P}\left(c_1^* = 1\right)\mathbb{P}\left(\tilde{X}_1 \notin [\frac{1}{2} - \left|\frac{1}{2} - X_{c_3^*}\right|, \frac{1}{2} + \left|\frac{1}{2} - X_{c_3^*}\right|]\right) \text{ by independence} \\
&= \frac{1}{n}\mathbb{E}\left[1 - 2\left|\frac{1}{2} - X_{c_3^*}\right|\right] = \frac{1}{n} - \frac{2}{n}\mathbb{E}\left[\left|\frac{1}{2} - X_{c_3^*}\right|\right]
\end{aligned}$$

where $c_3^* := \arg\min_{i \in [n]\setminus\{c_1^*\}} \left|X_i - \frac{1}{2}\right|$ is the index of the second nearest neighbor of $\frac{1}{2}$ among $X_{1:n}$. Note that c_3^* is not necessarily equal to c_2^*, since the definition of c_2^* requires $X_{c_2^*}$ to be on the other side of $\frac{1}{2}$ from $X_{c_1^*}$ while that of c_3^* does not.

By our knowledge of the expectation of the second order statistic among i.i.d. uniform random variables, we obtain that

$$\frac{1}{n} - \frac{2}{n}\mathbb{E}\left[\left|\frac{1}{2} - X_{c_3^*}\right|\right] = \frac{1}{n} - \frac{2}{n} \cdot \frac{1}{2} \cdot \frac{2}{n+1} = \frac{1}{n} - \frac{2}{n(n+1)}.$$

Therefore, similarly to before, we also have that there is a constant $\tilde{c}, c > 0$ such that

$$\|\nu(Z,\hat{g}) - \nu(Z,\hat{g}^{(-1)})\|_2 = \sqrt{n}\sqrt{\mathbb{P}\Big(X \in \mathcal{E}(X_{c_1^*}, X_{c_2^*}) \triangle \mathcal{E}\Big(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\Big)\Big)}$$

$$\geq \sqrt{n}\sqrt{\mathbb{P}\Big(X \in \mathcal{E}(X_{c_1^*}, X_{c_2^*}) \triangle \mathcal{E}\Big(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\Big)\Big| c_1^{*(-1)} \neq c_1^*\Big)\mathbb{P}\Big(c_1^{*(-1)} \neq c_1^*\Big)}$$

$$\geq \sqrt{n}\sqrt{\mathbb{P}\Big(X \in \mathcal{E}(X_{c_1^*}, X_{c_2^*}) \triangle \mathcal{E}\Big(X^{(-1)}_{c_1^{*(-1)}}, X^{(-1)}_{c_2^{*(-1)}}\Big)\Big| c_1^{*(-1)} \neq c_1^*, c_2^{*(-1)} = c_2^*\Big)\mathbb{P}\Big(c_1^{*(-1)} \neq c_1^*\Big)}$$

$$= \sqrt{n}\sqrt{\mathbb{E}\Big[\frac{1}{2}\Big|X_{c_1^*} - X^{(-1)}_{c_1^{*(-1)}}\Big|\Big| c_1^{*(-1)} \neq c_1^*, c_2^{*(-1)} = c_2^*\Big]\mathbb{P}\Big(c_1^{*(-1)} \neq c_1^*\Big)}$$

$$\geq \tilde{c}\sqrt{\mathbb{P}\Big(c_1^{*(-1)} \neq c_1^*\Big)} \geq \frac{c}{\sqrt{n}}.$$

Therefore we proved the second point. The third point follows because

$$\sqrt{n}\Big[V_n(\hat{g}) - V_n(g_0) - \big(V(\hat{g}) - V(g_0)\big)\Big] = -\sqrt{n}V(\hat{g}) = -\sqrt{n}V(\hat{g}) = \lambda_{X^*_{c_1^*,c_2^*}} \xrightarrow{d} \text{Exp}(0.5)$$

where $V_n(g_0) = V(g_0) = 0$ by definition, and $V_n(\hat{g})$ since the nearest neighbor estimator evaluated at a training data point never misclassifies the point. □

D Proof of Corollary 4

Proof. We note that by monotonicity of L^p norms, plugging the bound in (Algorithmic Stability) into the right hand side terms of (1) gives the stability conditions in lemma 2. Corollary 4 then immediately follows. □

E Proof of Theorem 5

Proof. Denote $Z^b_{1:m,(-l)}, b \in \{1, \dots, B\}$ as the corresponding bagged samples when the l-th data point Z_l is replaced with an independent copy $\tilde{Z}_l$. We have that for $l \in [n]$:

$$\begin{aligned}
&\left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} \\
&= \left\| \sup_x \left\| \frac{1}{B} \sum_{b=1}^{B} \left(\hat{h}(Z^b_{1:m})(x) - \hat{h}(Z^b_{1:m,(-l)})(x) \right) \right\|_2 \right\|_{2r} \\
&= \left\| \sup_x \left\| \frac{1}{B} \sum_{b=1}^{B} \left(\hat{h}(Z^b_{1:m})(x) - \hat{h}(Z^b_{1:m,(-l)})(x) \right) \mathbb{1}_{\{\exists t \leq m \text{ s.t. } Z^b_t = Z_l\}} \right\|_2 \right\|_{2r}.
\end{aligned}$$

The last equality was because

$$\hat{h}(Z^b_{1:m})(x) - \hat{\theta}(Z^b_{1:m,(-l)})(x) \neq 0$$

only if

$$\exists t \leq m \text{ s.t. } Z^b_t = Z_l.$$

Fix any $l \in [n]$. To simplify notations, let

$$\nabla\hat{h}(Z^b)(x) := \hat{h}(Z^b_{1:m})(x) - \hat{h}(Z^b_{1:m,(-l)})(x),$$

and let A_b be the event $\{\exists t \leq m \text{ s.t. } Z^b_t = Z_l\}$.

Then

$$\left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} = \left\| \sup_x \left\| \frac{1}{B} \sum_{b=1}^{B} \nabla\hat{h}(Z^b)(x) \mathbb{1}_{A_b} \right\|_2 \right\|_{2r}.$$

By triangle inequality and symmetry of distributions

$$
\begin{aligned}
&\left\|\sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2\right\|_{2r} = \left\|\sup_x \left\|\frac{1}{B}\sum_{b=1}^{B} \nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2\right\|_{2r} \\
&\leq \frac{1}{B}\left\|\sum_{b=1}^{B} \sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2\right\|_{2r} \\
&\leq \frac{1}{B}\left\|\sum_{b=1}^{B} \left\{\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 - \mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 \middle| Z_1, \ldots, Z_n, \tilde{Z}_l\right]\right\}\right\|_{2r} \\
&+ \frac{1}{B}\sum_{b=1}^{B}\left\|\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 \middle| Z_1, \ldots, Z_n, \tilde{Z}_l\right]\right\|_{2r} \\
&\leq \frac{1}{B}\left\|\sum_{b=1}^{B} \left\{\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 - \mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 \middle| Z_1, \ldots, Z_n, \tilde{Z}_l\right]\right\}\right\|_{2r} \\
&+ \left\|\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^1)(x)\mathbb{1}_{A_1}\right\|_2 \middle| Z_1, \ldots, Z_n, \tilde{Z}_l\right]\right\|_{2r}.
\end{aligned}
$$

For ease of notations, denote

$$Z_{(l)} := \left(Z_1, \ldots, Z_n, \tilde{Z}_l\right)$$

and denote

$$R_b := \sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 - \mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 \middle| Z_{(l)}\right].$$

For the first term, we have by tower law that

$$
\begin{aligned}
&\frac{1}{B}\left\|\sum_{b=1}^{B} \left\{\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 - \mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2 \middle| Z_{(l)}\right]\right\}\right\|_{2r} = \frac{1}{B}\left\|\sum_{b=1}^{B} R_b\right\|_{2r} \\
&= \frac{1}{B}\mathbb{E}\left[\left(\sum_{b=1}^{B} R_b\right)^{2r}\right]^{\frac{1}{2r}} = \frac{1}{B}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{b=1}^{B} R_b\right)^{2r} \middle| Z_{(l)}\right]\right]^{\frac{1}{2r}}.
\end{aligned}
$$

To further simplify, we use the following lemma:

Lemma 7 (Marcinkiewicz-Zygmund inequality)**.** *Let $p \geq 1$. If $X_1, \ldots, X_n$ are i.i.d. random variables such that $\mathbb{E}[X_1] = 0$, then there exists a constant C_p such that*

$$\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\|_p \leq C_p\|X_i\|_p.$$

Since $R_b, b \in [B]$ are i.i.d. conditional on $Z_{(l)}$, we then have that

$$
\begin{aligned}
&\frac{1}{B}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{b=1}^{B} R_b\right)^{2r} \middle| Z_{(l)}\right]\right]^{\frac{1}{2r}} \\
&\leq \frac{1}{B}\cdot\mathbb{E}\left[\left(\sqrt{B}C_{2r}\right)^{2r}\mathbb{E}\left[R_b^{2r}\middle| Z_{(l)}\right]\right]^{1/2r} = \frac{C_{2r}}{\sqrt{B}}\cdot\|R_b\|_{2r} \text{ by tower law.}
\end{aligned}
$$

Now also

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^1)(x)\mathbb{1}_{A_1}\right\|_2 \middle| Z_1,\ldots,Z_n,\tilde{Z}_l\right]\right\|_{2r} \\
&= \left\|\mathbb{P}\left(A_1\middle|Z_1,\ldots,Z_n,\tilde{Z}_l\right)\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^1)(x)\right\|_2 \middle| Z_1,\ldots,Z_n,\tilde{Z}_l,A_1\right]\right\|_{2r} \\
&\qquad\qquad \text{since } \nabla\hat{h}(Z^1)(x)=0 \text{ for all } x \text{ on } A_1^c \\
&= \left\|\mathbb{P}\left(A_1\right)\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^1)(x)\right\|_2 \middle| Z_1,\ldots,Z_n,\tilde{Z}_l,A_1\right]\right\|_{2r} \\
&\qquad\qquad \text{since } A_1 \text{ is independent of } Z_1,\ldots,Z_n,\tilde{Z}_l \\
&= \mathbb{P}\left(A_1\right)\left\|\mathbb{E}\left[\sup_x \left\|\nabla\hat{h}(Z^1)(x)\right\|_2 \middle| Z_1,\ldots,Z_n,\tilde{Z}_l,A_1\right]\right\|_{2r} \\
&\leq \mathbb{P}\left(A_1\right)\left\|\sup_x \left\|\nabla\hat{h}(Z^1)(x)\right\|_2\right\|_{2r} \text{ by Jensen's inequality and tower law} \\
&\leq \mathbb{P}\left(A_1\right)\cdot 2C \text{ by moment condition, since } 2r\leq s.
\end{aligned}
$$

By union bound,

$$\mathbb{P}\left(A_1\right)\leq\sum_{t=1}^m \mathbb{P}\left(Z_t^1=Z_l\right)=\frac{m}{n}.$$

Therefore, altogether we obtain

$$
\begin{aligned}
&\left\|\sup_x\|\hat{g}(x)-\hat{g}^{(-l)}(x)\|_2\right\|_{2r} \\
&\leq \frac{C_{2r}}{\sqrt{B}}\cdot\|R_b\|_{2r}+2C\cdot\frac{m}{n} \\
&\leq \frac{C_{2r}}{\sqrt{B}}\cdot\left\|\sup_x\left\|\nabla\hat{h}(Z^1)(x)\mathbb{1}_{A_1}\right\|_2\right\|_{2r}+\frac{C_{2r}}{\sqrt{B}}\cdot\left\|\mathbb{E}\left[\sup_x\left\|\nabla\hat{h}(Z^b)(x)\mathbb{1}_{A_b}\right\|_2\middle|Z_{(l)}\right]\right\|_{2r}+2C\cdot\frac{m}{n} \\
&\leq \frac{C_{2r}}{\sqrt{B}}\cdot\left\|\sup_x\left\|\nabla\hat{h}(Z^1)(x)\mathbb{1}_{A_1}\right\|_2\right\|_{2r}+\frac{C_{2r}}{\sqrt{B}}\cdot 2C\cdot\frac{m}{n}+2C\cdot\frac{m}{n} \\
&\leq \frac{2C\cdot C_{2r}}{\sqrt{B}}\cdot\left\|\mathbb{1}_{A_1}\right\|_k+\frac{C_{2r}}{\sqrt{B}}\cdot 2C\cdot\frac{m}{n}+2C\cdot\frac{m}{n} \\
&= \frac{2C\cdot C_{2r}}{\sqrt{B}}\cdot\left(\mathbb{P}(A_1)\right)^{1/k}+\frac{C_{2r}}{\sqrt{B}}\cdot 2C\cdot\frac{m}{n}+2C\cdot\frac{m}{n} \\
&\leq \frac{2C\cdot C_{2r}}{\sqrt{B}}\cdot\left(\frac{m}{n}\right)^{1/k}+\frac{C_{2r}}{\sqrt{B}}\cdot 2C\cdot\frac{m}{n}+2C\cdot\frac{m}{n}.
\end{aligned}
$$

Hence, we also have

$$\max_{l\leq n}\left\|\sup_x\|\hat{g}(x)-\hat{g}^{(-l)}(x)\|_2\right\|_{2r}\leq\frac{2C\cdot C_{2r}}{\sqrt{B}}\cdot\left(\frac{m}{n}\right)^{1/k}+\frac{C_{2r}}{\sqrt{B}}\cdot 2C\cdot\frac{m}{n}+2C\cdot\frac{m}{n}.$$

Assuming

$$m=o(\sqrt{n})$$

and

$$B>>m^{2/k}\cdot n^{1-\frac{2}{k}},$$

this upper bound is of order $o(n^{-1/2})$:

$$\max_{l\leq n}\left\|\sup_x\|\hat{g}(x)-\hat{g}^{(-l)}(x)\|_2\right\|_{2r}=o(n^{-1/2}).$$

□

Remark 3. *We could in fact relax the conditions in Theorem 5 by using Rosenthal's inequality instead of Marcinkiewicz-Zygmund inequality in the proof. Moreover, we can relax the bounded moments condition to restrict on L^{2r} norm instead of on L^s norm. This gives the following theorem.*

Theorem 8. *Assume B, m satisfy*

$$m = o(\sqrt{n}) \qquad B >> m^{\frac{1}{2r-1}} \cdot n^{\frac{r-1}{2r-1}},$$

and assume the base estimator $\hat{h}$ has bounded moments:

$$\max_{l \le n} \left\| \sup_x \left\| \hat{h}(Z^1_{1:m})(x) \right\|_2 \right\|_{2r} \le C$$

for some constant $C > 0$. Then (Algorithmic Stability) is achieved:

$$\max_{l \le n} \left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} = o(n^{-1/2}).$$

Therefore if a and ν satisfy the condition (1) then the condition (2) is satisfied.

Proof of Theorem 8. We follow the proof of Theorem 5 and obtain

$$\begin{aligned}
&\left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} = \left\| \sup_x \left\| \frac{1}{B} \sum_{b=1}^{B} \nabla \hat{h}(Z^b)(x) \mathbb{1}_{A_b} \right\|_2 \right\|_{2r} \\
&\le \frac{1}{B} \left\| \sum_{b=1}^{B} R_b \right\|_{2r} + 2C \cdot \frac{m}{n} \\
&= \frac{1}{B} \cdot \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{b=1}^{B} R_b \right)^{2r} \middle| Z_{(l)} \right] \right]^{1/2r} + 2C \cdot \frac{m}{n} \qquad \text{by tower law.}
\end{aligned}$$

We then use the following lemma.

Lemma 9 (Rosenthal's inequality)**.** *Let $p \ge 1$. If $X_1, \ldots, X_n$ are i.i.d. random variables such that $\mathbb{E}[X_1] = 0$, then there exists a constant C_p such that*

$$\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \|_p \le C_p \left(\|X_i\|_2 + n^{\frac{1}{p} - \frac{1}{2}} \|X_i\|_p \right).$$

Since $R_b, b \in [B]$ are i.i.d. conditional on $Z_{(l)}$, we then have that by symmetry of distributions

$$\begin{aligned}
&\frac{1}{B} \cdot \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{b=1}^{B} R_b \right)^{2r} \middle| Z_{(l)} \right] \right]^{1/2r} \\
&\le \frac{1}{B} \cdot \mathbb{E}\left[\left(\sqrt{B} C_{2r} \right)^{2r} \left(\left(\mathbb{E}\left[R_1^2 \middle| Z_{(l)} \right] \right)^{1/2} + B^{\frac{1}{2r} - \frac{1}{2}} \left(\mathbb{E}\left[R_1^{2r} \middle| Z_{(l)} \right] \right)^{1/2r} \right)^{2r} \right]^{1/2r} \\
&= \frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_1^2 \middle| Z_{(l)} \right] \right)^{1/2} + B^{\frac{1}{2r} - \frac{1}{2}} \left(\mathbb{E}\left[R_1^{2r} \middle| Z_{(l)} \right] \right)^{1/2r} \right\|_{2r}.
\end{aligned}$$

By triangle inequality, we have

$$\begin{aligned}
&\frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_1^2 \middle| Z_{(l)} \right] \right)^{1/2} + B^{\frac{1}{2r} - \frac{1}{2}} \left(\mathbb{E}\left[R_1^{2r} \middle| Z_{(l)} \right] \right)^{1/2r} \right\|_{2r} \\
&\le \frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_1^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} + \frac{C_{2r}}{\sqrt{B}} \cdot \left\| B^{\frac{1}{2r} - \frac{1}{2}} \left(\mathbb{E}\left[R_1^{2r} \middle| Z_{(l)} \right] \right)^{1/2r} \right\|_{2r}.
\end{aligned}$$

For further ease of notations let

$$R_{1,(1)} := \sup_x \left\| \nabla \hat{h}(Z^1)(x) \mathbb{1}_{A_1} \right\|_2,$$

$$R_{1,(2)} := \mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \mathbb{1}_{A_1} \right\|_2 \middle| Z_1, \ldots, Z_n, \tilde{Z}_l \right].$$

Note that

$$R_1 = R_{1,(1)} - R_{1,(2)}.$$

Then also by triangle inequality and Jensen's inequality

$$\begin{aligned}
&\frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_1^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} \\
&\leq \frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_{1,(1)}^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} + \frac{C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_{1,(2)}^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} \\
&\leq \frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[R_{1,(1)}^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} \\
&= \frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[\left(\sup_x \left\| \nabla \hat{h}(Z^1)(x) \mathbb{1}_{A_1} \right\|_2 \right)^2 \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} \\
&= \frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \cdot \mathbb{1}_{A_1} \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r}.
\end{aligned}$$

Further, we rewrite this term as

$$\begin{aligned}
&\frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \cdot \mathbb{1}_{A_1} \middle| Z_{(l)} \right] \right)^{1/2} \right\|_{2r} \\
&= \frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{P}\left(A_1 \middle| Z_{(l)} \right) \cdot \mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \middle| Z_{(l)}, A_1 \right] \right)^{1/2} \right\|_{2r} \\
&\qquad \text{since } \nabla \hat{h}(Z^1)(x) = 0 \text{ for all } x \text{ on } A_1^c \\
&= \frac{2C_{2r}}{\sqrt{B}} \cdot \left\| \left(\mathbb{P}(A_1) \cdot \mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \middle| Z_{(l)}, A_1 \right] \right)^{1/2} \right\|_{2r} \\
&\qquad \text{since } A_1 \text{ is independent of } Z_{(l)} \\
&= \frac{2C_{2r}}{\sqrt{B}} \cdot \left(\mathbb{P}(A_1) \right)^{1/2} \left\| \left(\mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \middle| Z_{(l)}, A_1 \right] \right)^{1/2} \right\|_{2r}.
\end{aligned}$$

By Jensen's inequality and tower law, we have

$$\begin{aligned}
&\frac{2C_{2r}}{\sqrt{B}} \cdot \left(\mathbb{P}(A_1) \right)^{1/2} \left\| \left(\mathbb{E}\left[\sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2^2 \middle| Z_{(l)}, A_1 \right] \right)^{1/2} \right\|_{2r} \\
&\leq \frac{2C_{2r}}{\sqrt{B}} \cdot \left(\mathbb{P}(A_1) \right)^{1/2} \left\| \sup_x \left\| \nabla \hat{h}(Z^1)(x) \right\|_2 \right\|_{2r} \\
&\leq \frac{2C_{2r}}{\sqrt{B}} \cdot \left(\frac{m}{n} \right)^{1/2} \cdot 2C.
\end{aligned}$$

Similarly, by replacing the powers of 2 and $1/2$ with $2r$ and $1/2r$, we can show that

$$\frac{C_{2r}}{\sqrt{B}} \cdot \left\| B^{\frac{1}{2r}-\frac{1}{2}} \left(\mathbb{E}\left[R_1^{2r} \middle| Z_{(l)} \right] \right)^{1/2r} \right\|_{2r}$$
$$\leq 2C_{2r} \cdot B^{\frac{1}{2r}-1} \left(\frac{m}{n} \right)^{1/2r} \cdot 2C.$$

Altogether, we have

$$\max_{l \leq n} \left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} = \left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r}$$
$$\leq \frac{2C_{2r}}{\sqrt{B}} \cdot \left(\frac{m}{n} \right)^{1/2} \cdot 2C + 2C_{2r} \cdot B^{\frac{1}{2r}-1} \left(\frac{m}{n} \right)^{1/2r} \cdot 2C + 2C \cdot \frac{m}{n}.$$

Assuming

$$m = o(\sqrt{n})$$

and

$$B >> m^{\frac{1}{2r-1}} \cdot n^{\frac{r-1}{2r-1}},$$

this upper bound is of order $o(n^{-1/2})$:

$$\max_{l \leq n} \left\| \sup_x \|\hat{g}(x) - \hat{g}^{(-l)}(x)\|_2 \right\|_{2r} = o(n^{-1/2}).$$

□

F Proof of Lemma 6

Proof. Lemma 6 follows from Theorem 3 of [35] by taking $\psi(Z;\theta) := Y - \theta(x)$. □

G Establishing L^{2r}- and Mean-Squared-Continuity for Examples 1, 2, and 3

Recall that

$$m(Z;\theta,g) = a(Z;g)\theta + \nu(Z;g).$$

We will establish L^{2r}-continuity and mean-squared-continuity for Examples 1, 2, and 3 in this section.

G.1 Establishing for Example 1

For this example, we have

$$a(Z;g) = (T - p(X))(T - p(X))'$$
$$\nu(Z;g) = (Y - q(X))(T - p(X))'.$$

Let $v \geq 1$ be the constant such that $1 = \frac{1}{r} + \frac{1}{v}$. Denote $T_i, i \in [p]$ as the i-th coordinate of T. Denote $Y_i, i \in [p]$ as the i-th coordinate of Y. For any function p denote $p_i(X), i \in [p]$ as the i-th coordinate of $p(X)$. We will show that subject to

$$\|T_i\|_{2v}, \|\hat{p}_i(X)\|_{2v}, \|\hat{p}_i(X_l)\|_{2v}, \|Y\|_{2v}, \|\hat{q}(X)\|_{2v}$$

being finite for all $i \in [p]$ then L^{2r}-continuity conditions hold.

Moreover, we will show that if further

$$\|T_i\|_{\infty}, \|p_i(X)\|_{\infty}, \|p_i'(X)\|_{\infty}, \|Y\|_{\infty}, \|q'(X)\|_{\infty}$$

are finite for all $i \in [p]$ then mean-squared-continuity conditions hold with $q = 2$.

We will illustrate for function a and for function ν separately.

For function a:
We first verify L^{2r}-continuity for function a. We have that for any $i, j \in [p]$

$$\begin{aligned}
&a_{i,j}(Z;\hat{g}) - a_{i,j}(Z;\hat{g}^{(-l)}) \\
&= \left(\hat{p}_i^{(-l)}(X) - \hat{p}_i(X)\right) T_j + T_i \left(\hat{p}_j^{(-l)}(X) - \hat{p}_j(X)\right) \\
&+ \frac{1}{2}\left(\hat{p}_i^{(-l)}(X) - \hat{p}_i(X)\right)\left(\hat{p}_j^{(-l)}(X) + \hat{p}_j(X)\right) \\
&+ \frac{1}{2}\left(\hat{p}_i^{(-l)}(X) + \hat{p}_i(X)\right)\left(\hat{p}_j^{(-l)}(X) - \hat{p}_j(X)\right).
\end{aligned}$$

Then by triangle inequality and Hölder's inequality, we obtain

$$\begin{aligned}
&\left\|a_{i,j}(Z;\hat{g}) - a_{i,j}(Z;\hat{g}^{(-l)})\right\|_2 \\
&\le \left\|\sup_x \left|\hat{p}_i^{(-l)}(x) - \hat{p}_i(x)\right| \cdot |T_j|\right\|_2 + \left\|\sup_x \left|\hat{p}_j^{(-l)}(x) - \hat{p}_j(x)\right| \cdot |T_i|\right\|_2 \\
&+ \frac{1}{2}\left\|\sup_x \left|\hat{p}_i^{(-l)}(x) - \hat{p}_i(x)\right| \cdot \left(\hat{p}_j^{(-l)}(X) + \hat{p}_j(X)\right)\right\|_2 \\
&+ \frac{1}{2}\left\|\sup_x \left|\hat{p}_j^{(-l)}(x) - \hat{p}_j(x)\right| \cdot \left(\hat{p}_i^{(-l)}(X) + \hat{p}_i(X)\right)\right\|_2 \\
&\le \left\|\sup_x \left|\hat{p}_i^{(-l)}(x) - \hat{p}_i(x)\right|\right\|_{2r} \cdot \|T_j\|_{2v} + \left\|\sup_x \left|\hat{p}_j^{(-l)}(x) - \hat{p}_j(x)\right|\right\|_{2r} \cdot \|T_i\|_{2v} \\
&+ \frac{1}{2}\left\|\sup_x \left|\hat{p}_i^{(-l)}(x) - \hat{p}_i(x)\right|\right\|_{2r} \cdot \left\|\hat{p}_j^{(-l)}(X) + \hat{p}_j(X)\right\|_{2v} \\
&+ \frac{1}{2}\left\|\sup_x \left|\hat{p}_j^{(-l)}(x) - \hat{p}_j(x)\right|\right\|_{2r} \cdot \left\|\hat{p}_i^{(-l)}(X) + \hat{p}_i(X)\right\|_{2v} \\
&\overset{(a)}{\le} L_1 \cdot \left\|\sup_x \left\|\hat{g}^{(-l)}(x) - \hat{g}(x)\right\|_2\right\|_{2r}
\end{aligned}$$

where

$$\begin{aligned}
L_1 &:= \|T_j\|_{2v} + \|T_i\|_{2v} + \frac{1}{2}\left\|\hat{p}_i^{(-l)}(X)\right\|_{2v} + \frac{1}{2}\|\hat{p}_i(X)\|_{2v} + \frac{1}{2}\left\|\hat{p}_j^{(-l)}(X)\right\|_{2v} + \frac{1}{2}\|\hat{p}_j(X)\|_{2v} \\
&\overset{(b)}{=} \|T_j\|_{2v} + \|T_i\|_{2v} + \|\hat{p}_i(X)\|_{2v} + \|\hat{p}_j(X)\|_{2v}.
\end{aligned}$$

Here for (a) we have used the triangle inequality, and for (b) we have used the fact that Z_l and $\tilde{Z}_l$ have the same distribution.

By replacing Z with Z_l, we can similarly show that

$$\begin{aligned}
&\left\|a_{i,j}(Z_l;\hat{g}) - a_{i,j}(Z_l;\hat{g}^{(-l)})\right\|_2 \\
&\le L_2 \cdot \left\|\sup_x \left\|\hat{g}^{(-l)}(x) - \hat{g}(x)\right\|_2\right\|_{2r}
\end{aligned}$$

where

$$L_2 := \|T_j\|_{2v} + \|T_i\|_{2v} + \frac{1}{2}\|\hat{p}_i(X)\|_{2v} + \frac{1}{2}\|\hat{p}_i(X_l)\|_{2v} + \frac{1}{2}\|\hat{p}_j(X)\|_{2v} + \frac{1}{2}\|\hat{p}_j(X_l)\|_{2v}.$$

Hence, the L^{2r}-continuity conditions hold for function a provided that all the aforementioned L^{2v}-norm quantities are finite.

Now we check the mean-squared-continuity conditions for function a.

For any g, g', any $i, j \in [p]$, we have

$$\begin{aligned}
&a_{i,j}(Z;g) - a_{i,j}(Z;g') \\
&= \left(p_i'(X) - p_i(X)\right) T_j + T_i \left(p_j'(X) - p_j(X)\right) \\
&+ \frac{1}{2}\left(p_i'(X) - p_i(X)\right)\left(p_j'(X) + p_j(X)\right) \\
&+ \frac{1}{2}\left(p_i'(X) + p_i(X)\right)\left(p_j'(X) - p_j(X)\right).
\end{aligned}$$

Then by triangle inequality and Hölder's inequality, we obtain

$$\begin{aligned}
&\|a_{i,j}(Z;g) - a_{i,j}(Z;g')\|_2 \\
&\le \|p_i'(X) - p_i(X)\|_2 \cdot \|T_j\|_\infty + \left\|p_j'(X) - p_j(X)\right\|_2 \cdot \|T_i\|_\infty \\
&+ \frac{1}{2}\|p_i'(X) - p_i(X)\|_2 \cdot \|p_j'(X) + p_j(X)\|_\infty + \frac{1}{2}\left\|p_j'(X) - p_j(X)\right\|_2 \cdot \|p_i'(X) + p_i(X)\|_\infty \\
&\le L_3 \cdot \|g - g'\|_2
\end{aligned}$$

where

$$L_3 := \|T_j\|_\infty + \|T_i\|_\infty + \frac{1}{2}\|p_j'(X)\|_\infty + \frac{1}{2}\|p_j(X)\|_\infty + \frac{1}{2}\|p_i'(X)\|_\infty + \frac{1}{2}\|p_i(X)\|_\infty.$$

Provided all these L^∞-norm quantities are finite, mean-squared-continuity conditions hold for function a with $q = 2$.

For function ν:
We have

$$\begin{aligned}
\nu_i(Z;\hat{g}) - \nu_i(Z;\hat{g}^{(-l)}) &= \left(\hat{q}^{(-l)}(X) - \hat{q}(X)\right) T_i + Y\left(\hat{p}_i^{(-l)}(X) - \hat{p}_i(X)\right) \\
&- \left(\hat{q}^{(-l)}(X) - \hat{q}(X)\right)\hat{p}_i(X) - \hat{q}^{(-l)}(X)\left(\hat{p}_i^{(-l)}(X) - \hat{p}_i(X)\right).
\end{aligned}$$

Hence, by triangle inequality and Hölder's inequality, we similarly obtain that

$$\begin{aligned}
&\left\|\nu_i(Z;\hat{g}) - \nu_i(Z;\hat{g}^{(-l)})\right\|_2 \\
&\le \left\|\sup_x \left|\hat{q}^{(-l)}(x) - \hat{q}(x)\right|\right\|_{2r} \cdot \{\|T_i\|_{2v} + \|\hat{p}_i(X)\|_{2v}\} \\
&+ \left\|\sup_x \left|\hat{p}_i^{(-l)}(x) - \hat{p}_i(x)\right|\right\|_{2r} \cdot \left\{\|Y\|_{2v} + \left\|\hat{q}^{(-l)}(X)\right\|_{2v}\right\} \\
&\le \left\|\sup_x \left\|\hat{g}^{(-l)}(x) - \hat{g}(x)\right\|_2\right\|_{2r} \cdot \{\|T_i\|_{2v} + \|\hat{p}_i(X)\|_{2v} + \|Y\|_{2v} + \|\hat{q}(X)\|_{2v}\}.
\end{aligned}$$

By replacing Z with Z_l, we can similarly show that

$$\begin{aligned}
&\left\|\nu_i(Z_l;\hat{g}) - \nu_i(Z_l;\hat{g}^{(-l)})\right\|_2 \\
&\le \left\|\sup_x \left\|\hat{g}^{(-l)}(x) - \hat{g}(x)\right\|_2\right\|_{2r} \cdot \{\|T_i\|_{2v} + \|\hat{p}_i(X_l)\|_{2v} + \|Y\|_{2v} + \|\hat{q}(X)\|_{2v}\}.
\end{aligned}$$

Therefore, the L^{2r}-continuity conditions hold for function ν provided that all these L^{2v}-norm quantities are finite.

As for mean-squared-continuity, note that we have that

$$\begin{aligned}
&\nu_i(Z;g) - \nu_i(Z;g') \\
&= \left(q'(X) - q(X)\right) T_i + Y\left(p_i'(X) - p_i(X)\right) \\
&- \left(q'(X) - q(X)\right) p_i(X) - q'(X)\left(p_i'(X) - p_i(X)\right).
\end{aligned}$$

Hence, by triangle inequality and Hölder's inequality, we derive that

$$\begin{aligned}
&\|\nu_i(Z;g) - \nu_i(Z;g')\|_2 \\
&\leq \|q'(X) - q(X)\|_2 \cdot \|T_i\|_\infty + \|p_i'(X) - p_i(X)\|_2 \cdot \|Y\|_\infty \\
&+ \|q'(X) - q(X)\|_2 \cdot \|p_i(X)\|_\infty + \|p_i'(X) - p_i(X)\|_2 \cdot \|q'(X)\|_\infty \\
&\leq \|g - g'\|_2 \cdot \{\|T_i\|_\infty + \|Y\|_\infty + \|p_i(X)\|_\infty + \|q'(X)\|_\infty\}.
\end{aligned}$$

Provided all these L^∞-norm quantities are finite, mean-squared-continuity conditions hold for function ν with $q = 2$.

G.2 Establishing for Example 2

For this example, we have

$$\begin{aligned}
a(Z;g) &= (Z - r(X))(T - p(X))' \\
\nu(Z;g) &= (Y - q(X))(Z - r(X))'.
\end{aligned}$$

Denote $Z_i, i \in [p]$ as the i-th coordinate of Z. Denote $r_i(X), i \in [p]$ as the i-th coordinate of $r(X)$. Analogously to Example 1, replacing all functions p_i and their estimates with r_i and their corresponding estimates, replacing all T_i with Z_i in the analysis of $a_{i,j}$ and ν_i, we can show that subject to

$$\|T_i\|_{2v}, \|Z_i\|_{2v}, \|\hat{p}_i(X)\|_{2v}, \|\hat{p}_i(X_l)\|_{2v}, \|\hat{r}_i(X)\|_{2v}, \|\hat{r}_i(X_l)\|_{2v}, \|Y\|_{2v}, \|\hat{q}(X)\|_{2v}$$

being finite for all $i \in [p]$ then L^{2r}-continuity conditions hold. Moreover, if further

$$\|T_i\|_\infty, \|Z_i\|_\infty, \|p_i(X)\|_\infty, \|p_i'(X)\|_\infty, \|r_i(X)\|_\infty, \|r_i'(X)\|_\infty, \|Y\|_\infty, \|q'(X)\|_\infty$$

are finite for all $i \in [p]$ then mean-squared-continuity conditions hold with $q = 2$.

G.3 Establishing for Example 3

For this example, we have

$$\begin{aligned}
a(Z;g) &\equiv 1 \\
\nu(Z;g) &= -m_b(Z;q) - \mu(T,X)(Y - q(T,X)).
\end{aligned}$$

The L^{2r}-continuity and mean-squared-continuity conditions trivially hold for function a. For function ν, we have

$$\begin{aligned}
&\nu(Z;\hat{g}) - \nu(Z;\hat{g}^{(-l)}) = (m_b(Z;\hat{q}^{(-l)}) - m_b(Z;\hat{q})) \\
&+ \hat{\mu}^{(-l)}(T,X)(\hat{q}(T,X) - \hat{q}^{(-l)}(T,X)) - (Y - \hat{q}(T,X))(\hat{\mu}(T,X) - \hat{\mu}^{(-l)}(T,X)).
\end{aligned}$$

Hence, by triangle inequality and Hölder's inequality we obtain

$$\begin{aligned}
&\|\nu(Z;\hat{g}) - \nu(Z;\hat{g}^{(-l)})\|_2 \\
&\leq \|m_b(Z;\hat{q}^{(-l)}) - m_b(Z;\hat{q})\|_2 \\
&+ \|\hat{\mu}^{(-l)}(T,X)(\hat{q}(T,X) - \hat{q}^{(-l)}(T,X))\|_2 + \|(Y - \hat{q}(T,X))(\hat{\mu}(T,X) - \hat{\mu}^{(-l)}(T,X))\|_2 \\
&\leq \|m_b(Z;\hat{q}^{(-l)}) - m_b(Z;\hat{q})\|_2 \\
&+ \|\hat{\mu}^{(-l)}(T,X)\|_{2v} \cdot \|\hat{q}(T,X) - \hat{q}^{(-l)}(T,X)\|_{2r} \\
&+ \|Y - \hat{q}(T,X)\|_{2v} \cdot \|\hat{\mu}(T,X) - \hat{\mu}^{(-l)}(T,X)\|_{2r} \\
&\leq \|m_b(Z;\hat{q}^{(-l)}) - m_b(Z;\hat{q})\|_2 \\
&+ \|\hat{\mu}(T,X)\|_{2v} \cdot \|\sup_{t,x} |\hat{q}(t,x) - \hat{q}^{(-l)}(t,x)|\|_{2r} \\
&+ (\|Y\|_{2v} + \|\hat{q}(T,X)\|_{2v})\|\sup_{t,x} |\hat{\mu}(t,x) - \hat{\mu}^{(-l)}(t,x)|\|_{2r}.
\end{aligned}$$

Since m_b is a linear functional, there exists $L_m > 0$ such that

$$\|m_b(Z;\hat{q}^{(-l)}) - m_b(Z;\hat{q})\|_2 \leq L_m \cdot \|\sup_{t,x} |\hat{q}(t,x) - \hat{q}^{(-l)}(t,x)|\|_{2r}.$$

Hence, we have

$$\begin{aligned}
&\|\nu(Z;\hat{g}) - \nu(Z;\hat{g}^{(-l)})\|_2 \\
&\leq \|\sup_{t,x} |\hat{g}(t,x) - \hat{g}^{(-l)}(t,x)|\|_{2r} \cdot \left\{L_m + \|\hat{\mu}(T,X)\|_{2v} + \|Y\|_{2v} + \|\hat{q}(T,X)\|_{2v}\right\}.
\end{aligned}$$

Analogously, by replacing Z with Z_l, we can show that

$$\begin{aligned}
&\|\nu(Z_l;\hat{g}) - \nu(Z_l;\hat{g}^{(-l)})\|_2 \\
&\leq \|\sup_{t,x} |\hat{g}(t,x) - \hat{g}^{(-l)}(t,x)|\|_{2r} \cdot \left\{\tilde{L}_m + \|\hat{\mu}(T,X)\|_{2v} + \|Y\|_{2v} + \|\hat{q}(T_l,X_l)\|_{2v}\right\}
\end{aligned}$$

for some constant $\tilde{L}_m > 0$.

Therefore, subject to

$$L_m, \tilde{L}_m, \|\hat{\mu}(T,X)\|_{2v}, \|Y\|_{2v}, \|\hat{q}(T,X)\|_{2v}, \|\hat{q}(T_l,X_l)\|_{2v}$$

being finite for all $i \in [p]$ then L^{2r}-continuity conditions also hold for function ν.

Moreover, we have

$$\begin{aligned}
&\|\nu(Z;g) - \nu(Z;g')\|_2 \\
&\leq \|m_b(Z;q') - m_b(Z;q)\|_2 \\
&+ \|\mu'(T,X)(q(T,X) - q'(T,X))\|_2 + \|(Y - q(T,X))(\mu(T,X) - \mu'(T,X))\|_2 \\
&\leq \|g - g'\|_2 \cdot \{L_b + \|\mu'(T,X)\|_\infty + \|Y\|_\infty + \|q(T,X)\|_\infty\}
\end{aligned}$$

where since m_b is a linear functional, there exists $L_b > 0$ such that

$$\|m_b(Z;q') - m_b(Z;q)\|_2 \leq L_b \cdot \|q(t,x) - q'(t,x)\|_2.$$

Provided that

$$L_b, \|\mu'(T,X)\|_\infty, \|Y\|_\infty, \|q(T,X)\|_\infty$$

are finite, mean-squared-continuity conditions hold for function ν with $q = 2$.

H Extension to Nonlinear Moments

In this section, we extend our results to the case where the moment function $m(Z;\theta,g)$ is not necessarily linear in the target parameter θ. For simplicity, we assume that the nuisance estimator $\hat{g}$ is symmetric in each of the training data points $Z_1, \ldots, Z_n$. Moreover, we will denote with Z a fresh random draw from the distribution.

We introduce some notation. We denote with $\|\cdot\|_{2,2}$ the norm of a random vector defined as: $\|Z\|_{2,2} = \sqrt{\mathbb{E}\left[\sum_i Z_i^2\right]}$, which can also be thought as taking the L_2 norm of each coordinate and and then taking the ℓ_2 vector norm of this vector, or equivalently taking the L_2 norm of the random variable defined as the ℓ_2 norm of the random vector. For clarity, for any random vector Z we will denote with $\|Z\|_{v,2}$ the random variable that corresponds to the ℓ_2 vector norm of the random vector, i.e. $\|Z\|_{v,2} = \sqrt{\sum_i Z_i^2}$. For any random variable V, denote with $\|V\|_2$ its ℓ_2 norm $\sqrt{\mathbb{E}[V^2]}$. Note that for any random vector Z, we have $\|Z\|_{2,2} = \|\|Z\|_{v,2}\|_2$.

Firstly, we establish a consistency lemma for $\hat{\theta}$. We note that in the linear moment case, such a separate proof of consistency was not required and a single step proof of asymptotic normality was feasible due to linearity. In the non-linear case, as is typical for moment based estimators, we first need to show that the estimate will eventually lie in a small ball around θ_0, and then argue normality. This is what the consistency lemma achieves.

Lemma 10 (Consistency)**.** *Assume that*

1. *The parameter space $\Theta \subset \mathbb{R}^p$ for target parameter θ is compact.*

2. *θ_0 is the unique solution of θ to the equation $M(\theta, g_0) = 0$.*

3. *The moment function $m(z; \theta, g)$ is uniformly continuous in θ over all Θ and a sufficciently small ℓ_2 ball $B_2(g_0) \subseteq \mathcal{G}$ around g_0. That is, $\forall \epsilon > 0$, $\exists \delta > 0$ such that for any $\tilde{\theta}_1, \tilde{\theta}_2 \in \Theta$ with $\|\tilde{\theta}_1 - \tilde{\theta}_2\|_2 < \delta$, $\forall g \in B_2(g_0)$, $\forall z$, we have*

$$\|m(z; \tilde{\theta}_1, g) - m(z; \tilde{\theta}_2, g)\|_2 < \epsilon.$$

4. *The moment function $m(Z; \theta, g)$ is mean-squared-continuous in g, uniformly in θ, i.e. $\exists L > 0$ and $q > 0$ such that:*

$$\max_{\theta \in \Theta} \|m(Z; \theta, g_1) - m(Z; \theta, g_2)\|_{2,2} \leq L \cdot \|g_1(Z) - g_2(Z)\|_{2,2}^q.$$

5. *Estimator $\hat{g}$ of the nuisance function is consistent: as $n \to \infty$*

$$\|\hat{g}(Z) - g_0(Z)\|_{2,2} = o(1).$$

6. *The moment function m and estimator $\hat{g}$ satisfies the following $o(1)$ leave-one-out stability condition:*

$$\max_{\theta \in \Theta} \|m(Z_1; \theta, \hat{g}) - m(Z_1; \theta, \hat{g}^{(-1)})\|_{2,2} = o(1)$$

as $n \to \infty$.

Then any estimator $\hat{\theta}$ that satisfies that $M_n(\hat{\theta}, \hat{g}) = o_p(1)$, also satisfies that $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. Fix any $\epsilon > 0$. Since $m(z; \theta, g)$ is uniformly continuous in $\theta \in \Theta$ for a sufficiently small ℓ_2-ball $B_2(g_0)$ around g_0, we have that, $\exists \delta > 0$ such that for any $\tilde{\theta}_1, \tilde{\theta}_2 \in \Theta$ with $\|\tilde{\theta}_1 - \tilde{\theta}_2\|_2 < \delta$, $\forall g \in B_2(g_0)$, $\forall z$, we have

$$\|m(z; \tilde{\theta}_1, g) - m(z; \tilde{\theta}_2, g)\|_2 < \epsilon/6.$$

Then

$$\|M(\tilde{\theta}_1, g) - M(\tilde{\theta}_2, g)\|_2 \leq \mathbb{E}\left[\|m(Z; \tilde{\theta}_1, g) - m(Z; \tilde{\theta}_2, g)\|_{v,2}\right] \leq \epsilon/6.$$

(that is, $M(\cdot, g_0)$ is uniformly continuous) and

$$\|M_n(\tilde{\theta}_1, g) - M_n(\tilde{\theta}_2, g)\|_{v,2} \leq \frac{1}{n} \sum_{i=1}^n \left\| m(Z_i; \tilde{\theta}_1, g) - m(Z_i; \tilde{\theta}_2, g) \right\|_{v,2} \leq \epsilon/6.$$

Since the parameter space Θ is compact, there exist $\theta_j, j = 1, \ldots, J$ such that

$$\Theta \subset \cup_{j=1}^J \mathcal{B}(\theta_j, \delta).$$

By Law of Large Numbers, we have $\forall j$, as $n \to \infty$

$$M_n(\theta_j, g_0) - M(\theta_j, g_0) \xrightarrow{p} 0.$$

Hence, $\forall \eta > 0$, for every j there exists n_j such that $\forall n > n_j$

$$\mathbb{P}\left(\|M_n(\theta_j, g_0) - M(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) < \frac{\eta}{3J}.$$

Then $\forall n > \max_j n_j$, we have

$$\mathbb{P}\left(\max_j \|M_n(\theta_j, g_0) - M(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) < \frac{\eta}{3}.$$

Moreover, for any $j \in J$, we have that:

$$\begin{aligned}
& \|M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0)\|_{2,2} \\
\leq\ & \max_\theta \|M_n(\theta, \hat{g}) - M_n(\theta, g_0)\|_{2,2} \\
=\ & \max_\theta \left\| \frac{1}{n} \sum_{i=1}^n \{m(Z_i; \theta, \hat{g}) - m(Z_i; \theta, g_0)\} \right\|_{2,2} \\
\leq\ & \frac{1}{n} \sum_{i=1}^n \max_\theta \|m(Z_i; \theta, \hat{g}) - m(Z_i; \theta, g_0)\|_{2,2} && \text{(triangle inequality)} \\
=\ & \max_\theta \|m(Z_1; \theta, \hat{g}) - m(Z_1; \theta, g_0)\|_{2,2} && \text{(symmetry of estimator)} \\
\leq\ & \max_\theta \left\| m(Z_1; \theta, \hat{g}) - m(Z_1; \theta, \hat{g}^{(-1)}) \right\|_{2,2} + \max_\theta \left\| m(Z_1; \theta, \hat{g}^{(-1)}) - m(Z_1; \theta, g_0) \right\|_{2,2} \\
\leq\ & \max_\theta \left\| m(Z_1; \theta, \hat{g}) - m(Z_1; \theta, \hat{g}^{(-1)}) \right\|_{2,2} + \max_\theta \left\| m(Z; \theta, \hat{g}^{(-1)}) - m(Z; \theta, g_0) \right\|_{2,2} \\
\leq\ & \max_\theta \left\| m(Z_1; \theta, \hat{g}) - m(Z_1; \theta, \hat{g}^{(-1)}) \right\|_{2,2} + L \cdot \left\| \hat{g}^{(-1)}(Z) - g_0(Z) \right\|_{2,2}^q \\
=\ & o(1).
\end{aligned}$$

Thus we have that $M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0) = o_p(1)$. Which means that $\forall \eta > 0$, there exists n_j such that for every $n > n_j$:

$$\mathbb{P}\left(\|M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) < \frac{\eta}{3J}.$$

Then $\forall n > \max_{j \in J} n_j$:

$$\mathbb{P}\left(\max_{j \in [J]} \|M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) < \frac{\eta}{3}.$$

Now $\forall \theta \in \Theta$, since $\Theta \subset \cup_{j=1}^J \mathcal{B}(\theta_j, \delta)$, there exists $k \in \{1, \ldots, J\}$ such that $\|\theta - \theta_k\|_2 < \delta$. Then for n sufficiently large, such that $\hat{g} \in B_2(g_0)$:

$$\begin{aligned}
& \|M_n(\theta, \hat{g}) - M(\theta, g_0)\|_{v,2} \\
\leq\ & \|M_n(\theta_k, \hat{g}) - M(\theta_k, g_0)\|_{v,2} + \|M_n(\theta, \hat{g}) - M_n(\theta_k, \hat{g})\|_{v,2} + \|M(\theta, g_0) - M(\theta_k, g_0)\|_{v,2} \\
\leq\ & \max_j \|M_n(\theta_j, \hat{g}) - M(\theta_j, g_0)\|_{v,2} + 2\epsilon/6.
\end{aligned}$$

Hence, we obtain

$$\mathbb{P}\left(\max_\theta \|M_n(\theta, \hat{g}) - M(\theta, g_0)\|_{v,2} > \epsilon \right) \leq\ \mathbb{P}\left(\max_j \|M_n(\theta_j, \hat{g}) - M(\theta_j, g_0)\|_{v,2} > \frac{2\epsilon}{3} \right).$$

Moreover, note that by the triangle inequality:

$$\begin{aligned}
& \max_j \|M_n(\theta_j, \hat{g}) - M(\theta_j, g_0)\|_{v,2} \\
\leq\ & \max_j \|M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0)\|_{v,2} + \max_j \|M_n(\theta_j, g_0) - M(\theta_j, g_0)\|_{v,2}.
\end{aligned}$$

Thus:

$$\begin{aligned}
& \mathbb{P}\left(\max_j \|M_n(\theta_j, \hat{g}) - M(\theta_j, g_0)\|_{v,2} > \frac{2\epsilon}{3} \right) \\
\leq\ & \mathbb{P}\left(\max_j \|M_n(\theta_j, \hat{g}) - M_n(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) + \mathbb{P}\left(\max_j \|M_n(\theta_j, g_0) - M(\theta_j, g_0)\|_{v,2} > \frac{\epsilon}{3} \right) \\
\leq\ & \frac{2\eta}{3} \leq \eta.
\end{aligned}$$

And we conclude that:

$$\mathbb{P}\left(\max_\theta \|M_n(\theta, \hat{g}) - M(\theta, g_0)\|_{v,2} > \epsilon \right) \leq\ \eta.$$

This shows that

$$\max_{\theta \in \Theta} \|M_n(\theta, \hat{g}) - M(\theta, g_0)\|_{v,2} \xrightarrow{p} 0$$

as $n \to \infty$.

In particular, this implies that

$$\|M_n(\hat{\theta}, \hat{g}) - M(\hat{\theta}, g_0)\|_{v,2} \xrightarrow{p} 0$$

as $n \to \infty$.

Hence, by triangle inequality and the fact that $M_n(\hat{\theta}, \hat{g}) = o_p(1)$, we obtain

$$\|M(\hat{\theta}, g_0)\|_{v,2} \leq \|M_n(\hat{\theta}, \hat{g})\|_{v,2} + \|M_n(\hat{\theta}, \hat{g}) - M(\hat{\theta}, g_0)\|_{v,2} = o_p(1).$$

Hence, $M(\hat{\theta}, g_0) = o_p(1)$.

It remains to show that

$$\hat{\theta} \xrightarrow{p} \theta_0$$

as $n \to \infty$.

To achieve this, again fix any $\epsilon > 0$. Then since Θ is compact, $B(\theta_0, \epsilon)^c$ is also compact as a closed subset of Θ. By continuity of $\theta \mapsto \|M(\theta, g_0)\|_2$ and the fact that θ_0 is the unique solution to $M(\theta_0, g_0) = 0$, we must have that $\|M(\theta, g_0)\|_{v,2}$ is bounded away from zero on $B(\theta_0, \epsilon)^c$. That is, $\exists \eta > 0$ such that for any θ with $\|\theta - \theta_0\|_{v,2} \geq \epsilon$,

$$\|M(\theta, g_0)\|_{v,2} > \eta.$$

Then since $M(\hat{\theta}, g_0) = o_p(1)$, there exists $N \in \mathbb{N}$ such that $\forall n > N$

$$\mathbb{P}\left(\|M(\hat{\theta}, g_0)\|_{v,2} > \eta\right) < \epsilon.$$

Then $\forall n > N$

$$\mathbb{P}\left(\|\hat{\theta} - \theta_0\|_{v,2} \geq \epsilon\right) \leq \mathbb{P}\left(\|M(\hat{\theta}, g_0)\|_{v,2} > \eta\right) < \epsilon.$$

This establishes consistency of $\hat{\theta}$. □

Now we extend Theorem 1 to nonlinear moments.

Theorem 11. *Let $A(\theta, g) := \partial_\theta M(\theta, g)$ denote the Jacobian of the moment vector, with respect to θ and $H_i(\theta, g) := \partial_\theta^2 M_i(\theta, g)$ denote the Hessian of the i-th moment coordinate. Suppose that the moment m is twice differentiable with $a(z; \theta, g) := \partial_\theta m(z; \theta, g)$ and $h_i(z; \theta, g) := \partial_\theta^2 m_i(z; \theta, g)$. Let $A_n(\theta, g) := \frac{1}{n} \sum_{i=1}^n a(Z_i; \theta, g)$ and $H_{i,n} := \frac{1}{n} \sum_{i=1}^n h(Z_i; \theta, g)$ denote the empirical counterparts of A, H_i.*

Suppose that the nuisance estimate $\hat{g} \in \mathcal{G}$ satisfies:

$$\|\hat{g} - g_0\|_2^2 \triangleq \mathbb{E}_X\left[\|\hat{g}(X) - g_0(X)\|_2^2\right] = o_p\left(n^{-1/2}\right). \qquad \text{(Consistency Rate)}$$

Suppose that the moment satisfies the Neyman orthogonality condition: for all $g \in \mathcal{G}$

$$D_g M(\theta_0, g_0)[g - g_0] \triangleq \frac{\partial}{\partial t} M(\theta_0, g_0 + t\,(g - g_0))\big|_{t=0} = 0 \qquad \text{(Neyman Orthogonality)}$$

and a second-order smoothness condition: for all $g \in \mathcal{G}$

$$D_{gg} M(\theta_0, g_0)[g - g_0] \triangleq \frac{\partial^2}{\partial t^2} M(\theta_0, g_0 + t\,(g - g_0))\big|_{t=0} = O\left(\|g - g_0\|_2^2\right) \qquad \text{(Smoothness)}$$

Assume that $A(\theta_0, g_0)^{-1}$ exists and that for any $g, g' \in \mathcal{G}$:

$$\|A(\theta_0, g) - A(\theta_0, g')\|_{op} = O\left(\|g - g'\|_2\right). \qquad \text{(Lipschitz)}$$

Suppose that the moment m and estimator $\hat{g}$ satisfy the stochastic equicontinuity conditions:

$$\begin{aligned} &\|A(\theta_0, \hat{g}) - A(\theta_0, g_0) - (A_n(\theta_0, \hat{g}) - A_n(\theta_0, g_0))\|_{op} = o_p(1) \\ &\sqrt{n}\,\|M(\theta_0, \hat{g}) - M(\theta_0, g_0) - (M_n(\theta_0, \hat{g}) - M_n(\theta_0, g_0))\|_{2,2} = o_p(1) \end{aligned} \quad \text{(NonLin. Stoc. Equi.)}$$

Suppose that the conditions in Lemma 10 are true. Moreover, assume that for any $i, j \in [p] \times [p]$, the random variable $a_{i,j}(Z; \theta_0, g_0)$ has bounded variance and that $\|\theta_0\|_2 = O(1)$. Suppose that $\exists$ open neighborhood W of g_0 such that

$$\sup_{\theta \in \Theta, g \in W, i \in [p]} \|H_i(\theta, g)\|_{op}, \|H_{i,n}(\theta, g)\|_{op} < \infty \quad \text{(Bounded Hessian)}$$

Let $\hat{\theta}$ denote any approximate solution to the plug-in empirical moment equation that satisfies $M_n(\hat{\theta}, \hat{g}) = o_p(n^{-1/2})$. Then $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{n \to \infty, d} N\left(0, A(\theta_0, g_0)^{-1} \mathbb{E}\left[m(Z; \theta_0, g_0)\, m(Z; \theta_0, g_0)^\top\right] A(\theta_0, g_0)^{-1}\right).$$

Proof. Let $A_i(\theta, g)$ denote the i-th column of the Jacobian matrix. By Taylor's expansion, we have $\forall g$ and $i \in [p]$

$$M_i(\hat{\theta}, g) - M_i(\theta_0, g) = A_i(\theta_0, g)'(\hat{\theta} - \theta_0) + (\hat{\theta} - \theta_0)' H_i(\tilde{\theta}_i, g)(\hat{\theta} - \theta_0)$$

for some $\tilde{\theta}_i$ between $\hat{\theta}$ and θ_0.

By Lemma 10, we know that

$$\|\hat{\theta} - \theta_0\|_2 = o_p(1).$$

Further, by the bounded Hessian condition and consistency of $\hat{\theta}$, we know that uniformly for $g \in W$

$$(\hat{\theta} - \theta_0)' H_i(\tilde{\theta}_i, g)(\hat{\theta} - \theta_0) = O_p\left(\|\hat{\theta} - \theta_0\|_2^2\right) = o_p(\|\hat{\theta} - \theta_0\|_2).$$

Moreover, for any g with $\|g - g_0\|_2 = o_p(1)$ we have

$$\begin{aligned} A(\theta_0, g) \cdot (\hat{\theta} - \theta_0) &= A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) + (A(\theta_0, g) - A(\theta_0, g_0)) \cdot (\hat{\theta} - \theta_0) \\ &= A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) + O(\|g - g_0\|_2 \, \|\hat{\theta} - \theta_0\|_2) \\ &= A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) + o_p(\|\hat{\theta} - \theta_0\|_2), \end{aligned}$$

where the second to last equality uses the Lipschitz condition.

By consistency, we have $\|\hat{g} - g_0\|_2 = o_p(1)$ and that $\hat{g} \in W$ for n large enough.

Hence, we obtain that for large enough n,

$$\begin{aligned} A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) &= M(\hat{\theta}, \hat{g}) - M(\theta_0, \hat{g}) + o_p(\|\hat{\theta} - \theta_0\|_2) \\ &= M(\hat{\theta}, \hat{g}) - M_n(\hat{\theta}, \hat{g}) + M(\theta_0, g_0) - M(\theta_0, \hat{g}) + M_n(\hat{\theta}, \hat{g}) + o_p(\|\hat{\theta} - \theta_0\|_2) \\ &= M(\hat{\theta}, \hat{g}) - M_n(\hat{\theta}, \hat{g}) + M(\theta_0, g_0) - M(\theta_0, \hat{g}) + o_p(n^{-1/2} + \|\hat{\theta} - \theta_0\|_2), \end{aligned}$$

where the last line follows because by definition $M_n(\hat{\theta}, \hat{g}) = o_p(n^{-1/2})$.

By Neyman orthogonality and the boundness condition on the second derivative of M with respect to g, for any g we have

$$M(\theta_0, g_0) - M(\theta_0, g) = D_g M(\theta_0, g_0)[g_0 - g] + O\left(\|g - g_0\|_2^2\right) = O\left(\|g - g_0\|_2^2\right).$$

Plugging in $g = \hat{g}$, noting that $\|g - g_0\|_2^2 = o_p(n^{-1/2})$ we obtain

$$M(\theta_0, g_0) - M(\theta_0, \hat{g}) = o_p(n^{-1/2}).$$

Let $G_n(\theta, g) := M(\theta, g) - M_n(\theta, g)$. Thus we have that

$$A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) = G_n(\hat{\theta}, \hat{g}) + o_p(n^{-1/2} + \|\hat{\theta} - \theta_0\|_2).$$

We decompose $G_n(\hat{\theta}, \hat{g})$ into the following sum:

$$G_n(\hat{\theta}, \hat{g}) = G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g}) + (G_n(\theta_0, \hat{g}) - G_n(\theta_0, g_0)) + G_n(\theta_0, g_0).$$

By Taylor's expansion, we have

$$\begin{aligned} G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g}) &= M(\hat{\theta}, \hat{g}) - M(\theta_0, \hat{g}) - \Big(M_n(\hat{\theta}, \hat{g}) - M_n(\theta_0, \hat{g})\Big) \\ &= \left(A(\theta_0, \hat{g}) - A_n(\theta_0, \hat{g})\right)' (\hat{\theta} - \theta_0) + o_p(\|\hat{\theta} - \theta_0\|_2) \end{aligned}$$

Now

$$\begin{aligned} & \|A(\theta_0, \hat{g}) - A_n(\theta_0, \hat{g})\|_{op} \\ \leq\ & \|A(\theta_0, g_0) - A_n(\theta_0, g_0)\|_{op} + \|A(\theta_0, \hat{g}) - A(\theta_0, g_0) - (A_n(\theta_0, \hat{g}) - A_n(\theta_0, g_0))\|_{op} \\ =\ & \|A(\theta_0, g_0) - A_n(\theta_0, g_0)\|_{op} + o_p(1), \end{aligned}$$

where the last equality follows from the stochastic equicontinuity condition on the Jacobian. Since $A(\theta_0, g_0) - A_n(\theta_0, g_0)$ is a mean zero empirical process with $\partial_\theta m(Z; \theta_0, g_0)$ having bounded variance, we have

$$\|A(\theta_0, g_0) - A_n(\theta_0, g_0)\|_{op} = o_p(1).$$

Hence,

$$G_n(\hat{\theta}, \hat{g}) - G_n(\theta_0, \hat{g}) = o_p(\|\hat{\theta} - \theta_0\|_2).$$

Moreover,

$$G_n(\theta_0, \hat{g}) - G_n(\theta_0, g_0) = \ M(\theta_0, \hat{g}) - M_n(\theta_0, \hat{g}) - (M(\theta_0, g_0) - M_n(\theta_0, g_0)) = o_p(n^{-1/2})$$

by stochastic equicontinuity.

In summary, we have

$$G_n(\hat{\theta}, \hat{g}) = G_n(\theta_0, g_0) + o_p(\|\hat{\theta} - \theta_0\|_2 + n^{-1/2}),$$

and thus

$$A(\theta_0, g_0) \cdot (\hat{\theta} - \theta_0) = G_n(\theta_0, g_0) + o_p(\|\hat{\theta} - \theta_0\|_2 + n^{-1/2}).$$

That is, we have

$$(\hat{\theta} - \theta_0) = A(\theta_0, g_0)^{-1} G_n(\theta_0, g_0) + o_p(\|\hat{\theta} - \theta_0\|_2 + n^{-1/2}).$$

Since $G_n(\theta_0, g_0)$ is a mean-zero empirical process, we have that $\|G_n(\theta_0, g_0)\|_2 = O_p(n^{-1/2})$.

By consistency of $\hat{\theta}$, above implies that

$$\hat{\theta} - \theta_0 = O_p(n^{-1/2}).$$

Thus we get

$$(\hat{\theta} - \theta_0) = A(\theta_0, g_0)^{-1} G_n(\theta_0, g_0) + o_p(n^{-1/2}).$$

By Slutsky's Theorem, we conclude that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, A(\theta_0, g_0)^{-1} \mathbb{E}[m(Z; \theta_0, g_0) m(Z; \theta_0, g_0)'] A(\theta_0, g_0)^{-1}\right).$$

□

We can also directly extend Lemma 2 to general nonlinear moments.

Lemma 12 (Non-Linear Main Lemma)**.** *If the estimation algorithm satisfies the stability conditions: for all* $i, j \in [p]$

$$\begin{aligned}
\max_{l\in[n]} \left\| a_{i,j}(Z_l;\theta_0,\hat{g}) - a_{i,j}(Z_l;\ \theta_0,\hat{g}^{(-l)}) \right\|_1 &= o(n^{-1/2}) \\
\max_{l\in[n]} \left\| a_{i,j}(Z;\theta_0,\hat{g}) - a_{i,j}(Z;\theta_0,\hat{g}^{(-l)}) \right\|_2 &= o(n^{-1/2}) \\
\max_{l\in[n]} \left\| m_i(Z_l;\theta_0,\hat{g}) - m_i(Z_l;\theta_0,\hat{g}^{(-l)}) \right\|_1 &= o(n^{-1/2}) \\
\max_{l\in[n]} \left\| m_i(Z;\theta_0,\hat{g}) - m_i(Z;\theta_0,g^{(-l)}) \right\|_2 &= o(n^{-1/2})
\end{aligned}$$

and the moment satisfies the mean-squared-continuity condition:

$$\begin{aligned}
\forall g, g' : \mathbb{E}[(a_{i,j}(Z;\theta_0,g) - a_{i,j}(Z;\theta_0,g'))^2] &\leq\ L\|g-g'\|_2^q \\
\forall g, g' : \mathbb{E}[(m_i(Z;\theta_0,g) - m_i(Z;\theta_0,g'))^2] &\leq\ L\|g-g'\|_2^q
\end{aligned}$$

for some $0 < q < \infty$ *and some* $L > 0$*, then the Condition* (NonLin. Stoc. Equi.) *is satisfied.*

Proof. The proof follows by replacing all functions $a(z,g), \nu(z,g)$ in the proof of Lemma 2 correspondingly with the functions $a(z;\theta_0,g)$ and $m(z;\theta_0,g)$. □

Application to bagging estimators. We finally note that if the moment satisfies Lipschitz conditions of the form:

$$\begin{aligned}
\mathbb{E}\left[\left(a_{i,j}(Z_l;\theta_0,\hat{g}) - a_{i,j}(Z_l;\theta_0,\hat{g}^{(-l)})\right)^2\right] &\leq\ L\cdot\mathbb{E}\Big[\sup_z \|\hat{g}(z)-\hat{g}^{(-l)}(z)\|_2^{2r}\Big]^{1/r} \\
\max_{\theta\in\Theta}\mathbb{E}\left[\left(m_i(Z_l;\theta,\hat{g}) - m_i(Z_l;\theta,\hat{g}^{(-l)})\right)^2\right] &\leq\ L\cdot\mathbb{E}\Big[\sup_z \|\hat{g}(z)-\hat{g}^{(-l)}(z)\|_2^{2r}\Big]^{1/r}
\end{aligned}$$

Then Theorem 5 can be applied to upper bound the right hand side of these inequalities by $o_p(n^{-1/2})$ for bagging estimators. This would then imply the stability conditions invoked in both the consistency and the normality theorem. Thus the main application for bagging estimators carries over to non-linear moments.